Tower: An Open Multilingual Large Language Model for Translation-Related Tasks

Year: 2024 | *Citations: 223* | *Authors: Duarte M. Alves, José P. Pombal, Nuno M. Guerreiro*

Abstract

While general-purpose large language models (LLMs) demonstrate proficiency on multiple tasks within the domain of translation, approaches based on open LLMs are competitive only when specializing on a single task. In this paper, we propose a recipe for tailoring LLMs to multiple tasks present in translation workflows. We perform continued pretraining on a multilingual mixture of monolingual and parallel data, creating TowerBase, followed by finetuning on instructions relevant for translation processes, creating TowerInstruct. Our final model surpasses open alternatives on several tasks relevant to translation workflows and is competitive with general-purpose closed LLMs. To facilitate future research, we release the Tower models, our specialization dataset, an evaluation framework for LLMs focusing on the translation ecosystem, and a collection of model generations, including ours, on our benchmark.